Miroslav Pospíšek

# Convergent algorithms suitable for the solution of the semiconductor device equations

*Applications of Mathematics*, Vol. 40 (1995), No. 2, 107–130

Persistent URL: http://dml.cz/dmlcz/134283

# CONVERGENT ALGORITHMS SUITABLE FOR THE SOLUTION OF THE SEMICONDUCTOR DEVICE EQUATIONS

MIROSLAV POSPÍŠEK, Praha

(Received February 14, 1992)

*Summary.* In this paper, two algorithms are proposed to solve systems of algebraic equations generated by a discretization procedure of the weak formulation of boundary value problems for systems of nonlinear elliptic equations. The first algorithm, *Newton-CG-MG*, is suitable for systems with gradient mappings, while the second, *Newton-CE-MG*, can be applied to more general systems. Convergence theorems are proved and application to the semiconductor device modelling is described.

*Keywords*: systems of nonlinear algebraic equations, semiconductor device equations

*AMS classification*: 35J65, 65H10

## 1. INTRODUCTION

In the preceding paper, [14], boundary value problems of the form (3.1) were studied. Conditions on the problem data, that are sufficient to define the weak formulation of the problem (3.1) and to guarantee existence of its weak solutions were shown there. Further, a discretization scheme based on numerical integration of the lower order terms only was examined and (weak) convergence of the discretized problem solutions to the weak solution of the problem (3.1) was proved.

To solve systems of nonlinear algebraic equations derived in the discretization procedure, two algorithms—*Newton-CG-MG* and *Newton-CE-MG*—are proposed in Section 4 of this paper. The algorithm *Newton-CG-MG* is suitable for the problems with gradient mappings and is based on the Newton method in conjuction with the method of conjugate gradients preconditioned by the variable V-cycle multigrid method. The algorithm *Newton-CE-MG* can be applied to more general problems. In this procedure, the Newton method is combined with the method of conjugate errors which is also preconditioned by the variable V-cycle multigrid method.

In Section 5, convergence theorems for both algorithms are proved. In both cases, the proof is based on the results of Bank and Rose [2] and recent developments of the multigrid theory [3].

In Section 6, the Van Roosbroeck's system (6.1)–(6.3) of three coupled nonlinear partial differential equations describing steady states of a semiconductor device is considered. It is shown that the algorithm *Newton-CE-MG* can be used for the solution of these problems and the algorithm *Newton-CG-MG* is also useful in some cases.

This paper together with [14] represents a method of treating boundary value problems in the form (3.1), starting with their weak formulation, up to convergent solution algorithm. As is shown, this approach can be applied to such a highly nonlinear system as the semiconductor device equations are. As far as the author knows, no similar approach to the semiconductor device equations resulting in a theoretically convergent multigrid based algorithm has been published yet. Besides, some more general results stated in Theorem 5.1 and Theorem 5.2 also seem to be new.

## 2. Basic notation

We will use the following notation:

| | |
|---|---|
| $\mathbb{N}$ | the set of non-negative integers, |
| $\mathbb{R}$ | the set of real numbers, |
| $\dot{\forall}$ | almost everywhere, |
| $\vec{n} = (n_1, \ldots, n_N)$ | vector of outward normal. |

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary divided into two disjoint subsets $\Gamma_D$ and $\Gamma_N$. Suppose that $\mu_1(\Gamma_D)$—the one-dimensional Lebesgue measure of $\Gamma_D$—is nonzero.

For a given vector function $u = (u_1, \ldots, u_m)$, $m \geqslant 1$, with sufficiently smooth components $u_i \colon \Omega \to \mathbb{R}$, $1 \leqslant i \leqslant m$, we denote

$$\nabla u = \left(\frac{\partial u_1}{\partial x_1}, \ldots, \frac{\partial u_m}{\partial x_1}, \frac{\partial u_1}{\partial x_2}, \ldots, \frac{\partial u_m}{\partial x_2}\right)$$

and

$$D^j u_i = \begin{cases} u_i, & j = 0, \\ \partial u_i / \partial x_j, & j = 1, 2. \end{cases}$$

If $\xi \in \mathbb{R}^{3m}$, we shall denote its components in the following way:

$$\xi = (\xi_{10}, \ldots, \xi_{m0}, \xi_{11}, \ldots, \xi_{m2}),$$

so that they correspond to the components of $(u, \nabla u)$.

Let $X$ be a real reflexive separable Banach space, equipped with the norm $\|\cdot\|_X$. The dual space of $X$ will be denoted by $X^*$ and the value of a continuous linear functional $F \in X^*$ on an element $v \in X$ will be denoted by

$$\langle F, v \rangle_X.$$

Let $H$ be a finite-dimensional Hilbert space with dimension $n$ and scalar product $\langle \cdot, \cdot \rangle_H$. Then, as is known, there exists a symmetric positive definite mapping $D\colon H \to H$ such that

$$(\forall u, v \in H)\ \big(\langle u, v \rangle_H = (Du, v) \equiv (u, Dv) \equiv (D^{1/2}u, D^{1/2}v)\big),$$

where $(\cdot, \cdot)$ denotes the Euclidean scalar product $(u, v) = \sum_{i=1}^{n} u_i v_i$. We shall often write $(u, v)_D$ and $\|u\|_D$, instead of $\langle u, v \rangle_H$ and $\|u\|_H$, respectively, and denote the space $H$ by $H_D$. For a mapping $A\colon H_D \to H_D$, the norm $\|A\|_D$ defined by

$$\|A\|_D = \sup_{\|v\|_D = 1} (Av, v)_D$$

will be used. In case $D \equiv I$, the indices in $(\cdot, \cdot)_D$ and $\|\cdot\|_D$ will be often omitted.

We also introduce here an abstract function space $V$, which will be referred to throughout the paper:

Let $1 < p < \infty$. The closure of the set

$$\{v \in C^\infty(\overline{\Omega})\colon v = 0 \text{ on } \Gamma_D\}$$

in the norm of $W_0^{1,p}(\Omega)$[1] will be denoted by $V^p$. The space $V$ is defined by

$$V = \prod_{i=1}^{m} V^{p_i}, \quad 1 < p_i < \infty,\ 1 \leqslant i \leqslant m, \tag{2.1}$$

and equipped with the norm

$$\|v\|_V = \Big(\sum_{i=1}^{m} \|v_i\|_{V^{p_i}}^{p_{\min}}\Big)^{\frac{1}{p_{\min}}} = \Bigg(\sum_{i=1}^{m}\Big(\sum_{j=1}^{N}\int_\Omega |D^j v_i|^{p_i}\,\mathrm{d}x\Big)^{\frac{p_{\min}}{p_i}}\Bigg)^{\frac{1}{p_{\min}}}, \tag{2.2}$$

where $p_{\min} = \min\{p_1, \ldots, p_m\}$.

---

[1] Recall that this norm can be defined as follows:

$$(\forall u \in W_0^{1,p}(\Omega))(\|u\|_{W_0^{1,p}(\Omega)} = \Big(\sum_{j=1}^{N} \int_\Omega |D^j u|^p\,\mathrm{d}x\Big)^{1/p}).$$

## 3. Problem formulation

Let $m$ and $\Omega$ be as in Section 2 and let functions

$$\begin{aligned} a_{ij}\colon &\ \Omega\times\mathbb{R}^{3m}\to\mathbb{R}, && i=1,\ldots,m, j=0,1,2,\\ f_i\colon &\ \Omega\to\mathbb{R}, && i=1,\ldots,m,\\ d_i\colon &\ \Omega\cup\Gamma_D\to\mathbb{R}, && i=1,\ldots,m,\\ h_i\colon &\ \Gamma_N\to\mathbb{R}, && i=1,\ldots,m \end{aligned}$$

be given. As in the preceding paper (Pospíšek [14]), we are interested in boundary value problems in the following form:

$$\begin{aligned} -\sum_{j=1}^{2} D^j a_{ij}(x;u,\nabla u)+a_{i0}(x;u) &= f_i, \quad i=1,\ldots,m,\ x\in\Omega,\\ u_i &= d_i, \quad i=1,\ldots,m,\ x\in\Gamma_D,\\ \sum_{j=1}^{2} n_j a_{ij}(x;u,\nabla u) &= h_i, \quad i=1,\ldots,m,\ x\in\Gamma_N. \end{aligned}\tag{3.1}$$

We recall that (precise meaning of the following conditions is given e.g. in Pospíšek [14]) if the functions $a_{ij}$ satisfy

(A1) Carathéodory conditions,

(A2) growth conditions with some coefficients $p_i>1$, $1\leqslant i\leqslant m$,

(A3) coercivity condition with the same coefficients as in (A2),

(A4) condition of strict monotonicity in principal part,

if the functions $d_i$, $f_i$ and $h_i$ have the properties

$$(D1)\ d_i\in W^{1,p_i}(\Omega),\ f_i\in L_{q_i}(\Omega),\ h_i\in L_{q_i}(\Gamma_N),\ 1/p_i+1/q_i=1,\ 1\leqslant i\leqslant m,$$

if the space $V$ is defined as in Section 2 with $p_i$, $i=1,\ldots,m$, from (A2) and if a mapping $A\colon V\to V^*$ and a functional $F\in V^*$ are defined by (here and in the sequel we denote $d=(d_1,\ldots,d_m)$)

$$(\forall u,v\in V)\Big(\langle Au,v\rangle_V=\sum_{i=1}^{m}\sum_{j=0}^{2}\int_\Omega a_{ij}(x;u+d,\nabla(u+d))D^j v_i\,\mathrm{d}x\Big),\tag{3.2}$$

$$(\forall v\in V)\Big(\langle F,v\rangle_V=\sum_{i=1}^{m}\Big(\int_\Omega f_i v_i\,\mathrm{d}x+\int_{\Gamma_N} h_i v_i\,\mathrm{d}S\Big)\Big),\tag{3.3}$$

then the problem

$$\text{(3.4)} \qquad \begin{aligned} &\text{Find } u \in V \text{ such that} \\ &(\forall v \in V)(\langle Au, v\rangle_V = \langle F, v\rangle_V), \end{aligned}$$

i.e. the weak formulation of the problem (3.1), can be formulated and has a solution. If, moreover, the strict monotonicity condition

$$\text{(3.5)} \qquad \begin{aligned} &(\forall x \in \Omega)(\forall \xi, \eta \in \mathbb{R}^{3m}, \xi \neq \eta) \\ &\Big(\sum_{i=1}^{m}\sum_{j=0}^{2}[a_{ij}(x;\xi) - a_{ij}(x;\eta)](\xi_{ij} - \eta_{ij}) > 0\Big) \end{aligned}$$

is fulfilled, the solution is unique.

Let $T_0$ be any conforming triangulation of $\Omega$ that is of weakly acute type—i.e. no internal angle of any triangle in $T_0$ is greater than $\pi/2$. We proceed as in Pospíšek [14]:

- Choose an integer $J \geqslant 0$.
- If $J > 0$, we refine $T_0$ by dividing each triangle $t \in T_0$ into four congruent triangles and thus obtain grid $T_1$. Applying the same procedure to the currently finest grid, we continue until the grid $T_J$ is generated.
- We construct dual meshes $B_j$, $0 \leqslant j \leqslant J$, by joining the midpoints of the edges with the centre of gravity in each triangle $t \in T_j$, $0 \leqslant j \leqslant J$. With each vertex $P \in T_l$ we associate a region $\omega_P$ consisting of those triangles $t \in T_l$ which have $P$ as a vertex and the so-called box $b_P \in B_l$, $b_P \subset \omega_P$, which consists of the union of the subregions in $\omega_P$ which again have $P$ as a vertex.

For further purposes, we denote

$$\begin{aligned} \Omega_j &= \{P \in \Omega \dot{-} \overline{\Gamma_D}, P \text{ is a vertex of } t \in T_j\}, \\ N_j &= \operatorname{card} \Omega_j, \end{aligned}$$

for $j = 0, \ldots, J$.

Now we use the finest grid $(T_J, B_J)$ to define the space $V_J$, a mapping $A_J \colon V_J \to V_J^*$ and a functional $F_J \in V_J^*$:

$$\begin{aligned} V_J = \{&v \in V, v \equiv (v_1, \ldots, v_m) \colon (\forall i, i = 1, \ldots, m)(\forall t \in T_J) \\ &(v_i \in C(\Omega)) \wedge (v_i|_t \text{ is linear})\}, \end{aligned}$$

$$\langle A_J u, v\rangle_V = \sum_{i=1}^{m}\sum_{j=1}^{2}\int_\Omega a_{ij}(x; u+d, \nabla(u+d)) D^j v_i \,dx$$

(3.6) $$+\sum_{i=1}^{m}\sum_{P\in\Omega_J} \mu_2(b_P) a_{i0}(P; (u+d)(P)) v_i(P),$$

(3.7) $$\langle F_J, v\rangle_V = \sum_{i=1}^{m}\sum_{P\in\Omega_J} (\mu_2(b_P) f_i(P) v_i(P) + \mu_1(b_P\cap\Gamma_N) h_i(P) v_i(P)).$$

It was shown in Pospíšek [14] that if the functions $a_{ij}$ satisfy conditions (A1)–(A4) and, moreover,

(A5) for all $i$, $i = 1,\ldots,m$, $a_{i0} \in C(\overline{\Omega}\times\mathbb{R}^m)$,

(D2) $f_i \in C(\overline{\Omega})$, $d_i \in C^1(\overline{\Omega})$, $h_i \in C(\Gamma_N)$, $i = 1,\ldots,m$,

then the problem

(3.8)
Find $u^J \in V_J$ such that
$(\forall v \in V_J)(\langle A_J u^J, v\rangle_V = \langle F_J, v\rangle_V)$

has a solution which, as $J \to \infty$, weakly converges to a solution of the problem (3.2)–(3.4).

Let $\prec$ be a complete ordering of the set $\Omega_j$. Define a mapping $\nu_j\colon \{1,2,\ldots,N_j\} \to \Omega_j$ such that

(3.9) $$(\forall k_1, k_2,\ 1 \leqslant k_1, k_2 \leqslant N_j)\ (k_1 < k_2 \Leftrightarrow \nu_j(k_1) \prec \nu_j(k_2)).$$

Clearly (see e.g. Pospíšek [14]), the problem (3.8) is equivalent to a system of (nonlinear) algebraic equations

(3.10) $$g(u^H) = 0 \quad \text{in } \mathbb{R}^{mN_J},$$

where $u^H \in \mathbb{R}^{mN_J}$ can be viewed as consisting of $m$ vectors $u_i^H \in \mathbb{R}^{N_J}$, $1 \leqslant i \leqslant m$,

$$u^H = (u_1^H, \ldots, u_m^H),$$

with each $u_i^H$ corresponding to the nodal values of $u_i^J$, the $i$-th component of $u^J$ from (3.8),

$$(\forall i, 1 \leqslant i \leqslant m)(\forall k, 1 \leqslant k \leqslant N_J)((u_i^H)_k = u_i^J(\nu_J(k))).$$

In this paper we describe algorithms suitable for the solution of the problem (3.10).

For this purpose we divide the set of equations in (3.10) into $m$ blocks so that each block corresponds to a discretization of one partial differential equation in (3.1). Then the system (3.10) can be written in the form

$$
\begin{aligned}
g_1(u_1^H, \ldots, u_m^H) &= 0 \\
g_2(u_1^H, \ldots, u_m^H) &= 0 \\
\ldots\ldots\ldots\ldots & \\
g_m(u_1^H, \ldots, u_m^H) &= 0
\end{aligned}
\tag{3.11}
$$

where $g_i\colon \mathbb{R}^{mN_J} \to \mathbb{R}^{N_J}$, $1 \leqslant i \leqslant m$. The value of the Fréchet derivative of $g(v)$ with respect to $v = (v_1, \ldots, v_m)$, $v_i \in \mathbb{R}^{N_J}$, $1 \leqslant i \leqslant m$, at a point $v_0$ can be expressed in the block form

$$
g'(v_0) \equiv \frac{\partial g(v_0)}{\partial v} = \left(\frac{\partial g_i(v_0)}{\partial v_j}\right)_{i,j=1,\ldots,m} \equiv (g'_{ij}(v_0))_{i,j=1,\ldots,m}. \tag{3.12}
$$

In the following we shall also use the fact that the mappings $g'(v_0)$ and $g'_{ii}(v_0)$ can be understood as discretizations (by the method (3.6)–(3.7)) of some linear mappings

$$
\mathcal{L}_0(v_0)\colon V_J \to V_J \qquad \text{and} \qquad \mathcal{L}_i(v_0)\colon V^{p_i} \to V^{p_i}, \tag{3.13}
$$

respectively.

## 4. Solution algorithms

We will describe two algorithms—*Newton-CG-MG* and *Newton-CE-MG*. In both algorithms, the overall strategy is the same:

- Modified Newton's method is used.
- Systems of linear equations arising in this method are solved by some kind of the conjugate direction method—the conjugate gradient method and the conjugate error method (see e.g. Samarskij, Nikolajev [17, sec. 8.3]) in the case of *Newton-CG-MG* and *Newton-CE-MG*, respectively.
- As a preconditioner of the conjugate direction method, the variable V-cycle multigrid method is used.

Now, we describe the individual parts of the algorithms:

**Procedure Newton.**

Input: $g\colon \mathbb{R}^n \to \mathbb{R}$, where $n \in \mathbb{N}$.

Output: $k \in \mathbb{N}$, $u_k \in \mathbb{R}^n$, where $u_k$ is an approximate solution of the equation $g(x) = 0$.

(N1) Choose initial approximation $u_0$ and $\delta \in (0,1)$.

(N2) Let $\mathcal{K} \in \mathbb{R}$, $\mathcal{K} \geqslant 0$. Set $\mathcal{K} := 0$, $\quad k := 0$, compute $g(u_0)$, $\|g(u_0)\|$.

(N3) Choose $\varrho \in \mathbb{N}$, $\varrho \geqslant 1$.
In the *Newton-CG-MG* method, $v_k := L^{CG}(g'(u_k), -g(u_k), \varrho)$.
In the *Newton-CE-MG* method, $v_k := L^{CE}(g'(u_k), -g(u_k), \varrho)$.

(The mappings $L^{CG}$ and $L^{CE}$ will be defined later, see PROCEDURE CG/CE.)

(N4) $\tau_k := (1 + \mathcal{K}\|g(u_k)\|)^{-1}$.

(N5) $u_{k+1} := u_k + \tau_k v_k$, compute $g(u_{k+1})$, $\|g(u_{k+1})\|$.

(N6) **if** $(1 - \|g(u_{k+1})\|/\|g(u_k)\|) < \tau_k \delta$ **then**
increase $\mathcal{K}$, **go to** (N4),
**else**
decrease $\mathcal{K}$, $\quad k := k + 1$,
**endif**

(N7) **if** (convergence) **then**
**exit else**
**go to** (N3)
**endif**

In step (N3), the mappings $L^{CG}$ and $L^{CE}$ are defined by several steps of the conjugate gradient method and the conjugate error method, respectively. As is known, both methods are special cases of the conjugate direction method and thus we will describe them both in one procedure.

**Procedure CG/CE.**

Input: $A\colon \mathbb{R}^n \to \mathbb{R}^n$, $b \in \mathbb{R}^n$, $n \in \mathbb{N}$, $\varrho \in \mathbb{N}$

Output: $x_{\varrho+1}$, an approximate solution of linear algebraic system $Ax = b$, also denoted as

- $L^{CG}(A, b, \varrho)$ in the CG method,
- $L^{CE}(A, b, \varrho)$ in the CE method.

(CD1) Choose initial approximation $x_0 = 0$, set $r_0 := -b \equiv Ax_0 - b$.

(CD2) In the CG method, $w_0 := B^{CG} r_0$, $\sigma_1 := (r_0, w_0)/(Aw_0, w_0)$.
In the CE method, $w_0 := B^{CE} r_0$, $\sigma_1 := (r_0, r_0)/(Aw_0, r_0)$.
(The mappings $B^{CG}$ and $B^{CE}$ will be defined later, see (4.8), (4.9).)

(CD3) $x_1 := x_0 - \sigma_1 w_0$.

(CD4) **For** $k = 1, 2, \ldots, \varrho$,
$r_k := Ax_k - b$.

In the CG method,

$$w_k := B^{CG} r_k, \quad \sigma_{k+1} = \frac{(r_k,\ w_k)}{(Aw_k,\ w_k)},$$
$$\alpha_{k+1} = \Big(1 - \frac{\sigma_{k+1}}{\sigma_k} \frac{(r_k,\ w_k)}{(r_{k-1},\ w_{k-1})} \frac{1}{\alpha_k}\Big)^{-1}.$$

In the CE method,

$$w_k := B^{CE} r_k, \quad \sigma_{k+1} = \frac{(r_k,\ r_k)}{(Aw_k,\ r_k)},$$
$$\alpha_{k+1} = \Big(1 - \frac{\sigma_{k+1}}{\sigma_k} \frac{(r_k,\ r_k)}{(r_{k-1},\ r_{k-1})} \frac{1}{\alpha_k}\Big)^{-1}.$$

In both cases:

$$x_{k+1} := \alpha_{k+1}(x_k - \sigma_k w_k) + (1 - \alpha_{k+1}) x_{k-1}.$$

We will often write $A_k$ instead of $g'(u_k)$ and $L_k^{CG} b$ and $L_k^{CE} b$ instead of $L^{CG}(A_k, b, \varrho)$ and $L^{CE}(A_k, b, \varrho)$, respectively.

In the above procedure, the mappings $B^{CG}$ and $B^{CE}$ representing the so-called preconditioning remain to be defined. In both cases, this is done by means of the so-called variable V-cycle multigrid method, see e.g. Bramble, Pasciak, Xu [3]. Hence, before specifying those mappings, we shall describe briefly the multigrid method.

**Procedure MG.**

Input:

- An integer $J \geqslant 0$.
- Finite-dimensional Hilbert spaces $H_j$, $j = 0, 1, \ldots, J$, with the scalar product in $H_j$ denoted by $(\cdot, \cdot)_j$.
- Symmetric positive definite mappings

$$\mathcal{A}_j \colon H_j \to H_j, \quad j = 0, 1, \ldots, J. \tag{4.1}$$

- Linear mappings

$$\mathcal{I}_j \colon H_{j-1} \to H_j, \quad j = 1, \ldots, J.$$

- Mappings

$$\mathcal{P}_{j-1} \colon H_j \to H_{j-1}, \quad \mathcal{P}^o_{j-1} \colon H_j \to H_{j-1}, \quad j = 1, \ldots, J,$$

defined for $j = 1, \dots, J$ by $(\forall\psi \in H_j)(\forall\varphi \in H_{j-1})$

$$(\mathcal{A}_{j-1}\mathcal{P}_{j-1}\psi, \varphi)_{j-1} = (\mathcal{A}_j\psi, \mathcal{I}_j\varphi)_j, \qquad (\mathcal{P}^o_{j-1}\psi, \varphi)_{j-1} = (\psi, \mathcal{I}_j\varphi)_j.$$

- Linear mappings $\mathcal{R}_j \colon H_j \to H_j$, $j = 1, \dots, J$.
- Integers $n(j), j = 0, \dots, J$, such that

$$(4.2) \qquad (\exists\beta_0 > 1)(\exists\beta_1 \geqslant \beta_0)(\forall j, j = 1, \dots, J) \quad (\beta_0 n(j) \leqslant n(j-1) \leqslant \beta_1 n(j))$$

is valid.

Output: Mappings $M_j \colon H_j \to H_j$, $j = 0, \dots, J$.

Mappings are defined by induction. Set $M_0 := A_0^{-1}$. Assume that $0 < j \leqslant J$, $M_{j-1}$ has been defined and $f \in H_j$, $y^0, \dots, y^{2n(j)} \in H_j$. We define $M_j f$ as follows:

(MG1) $y^0 := 0$.

(MG2) **for** $l = 1, \dots, n(j)$

$$(4.3) \qquad y^l := y^{l-1} + \mathcal{R}_j(f - \mathcal{A}_j y^{l-1}).$$

(MG3) $y^{n(j)} := y^{n(j)} + \mathcal{I}_j q$, where

$$(4.4) \qquad q := M_{j-1}[\mathcal{P}^o_{j-1}(f - \mathcal{A}_j y^{n(j)})].$$

(MG4) **for** $l = n(j)+1, \dots, 2n(j)$

$$(4.5) \qquad y^l := y^{l-1} + \mathcal{R}_j(f - \mathcal{A}_j y^{l-1}).$$

(MG5) $M_j f := y^{2n(j)}$.

Having described the multigrid method, we shall now specify the mappings $B^{CG}$ and $B^{CE}$ from the PROCEDURE CG/CE.

In the $k$-th Newton step of the algorithm Newton-CG-MG we take:

- $J$ from the discretization procedure, see e.g. (3.8).
- $H_j = \mathbb{R}^{mN_j}$, $0 \leqslant j \leqslant J$, with the scalar product

$$(4.6) \qquad (\forall u, v \in H_j)\left((u,v)_{H_j} = \sum_{l=1}^{N_j} \mu_2(b_{\nu_j(l)}) \sum_{i=1}^{m} u_{N_j(i-1)+l}\, v_{N_j(i-1)+l}\right)$$

where $\nu_j$ is the mapping from (3.9).

- Mappings $\mathcal{A}_j$, $j = 0, \dots, J$, defined as the discretization of the mapping $\mathcal{L}_0(u_k)$ on the grids $(T_j, B_j)$ by the method (3.6)–(3.7), with $u_k$ from the PROCEDURE

NEWTON and $\mathcal{L}_0$ defined as in (3.13). (Here we must ensure symmetry and positive definiteness of $\mathcal{A}_j$.)

- $\mathcal{I}_j$, $j = 1, \ldots, J$, as linear interpolations from grid $T_{j-1}$ to $T_j$.
- $\mathcal{R}_j$ such that they correspond to one sweep of the symmetric Gauss-Seidel method: if we write $\mathcal{A}_j$ from (4.1) in the form $\mathcal{A}_j = L_j + D_j + L_j^T$, where $L_j$ is a strictly lower triangular and $D_j$ a diagonal matrix, we set

$$\mathcal{R}_j \equiv (L_j^T + D_j)^{-1} D_j (L_j + D_j)^{-1} \tag{4.7}$$

- $n(J) = 1$, $n(j) = 2n(j+1)$, $j = 0, \ldots, J-1$.

We denote the mapping from the PROCEDURE MG with the above settings by $M_J^{(0,k)}$ and put

$$B^{CG} = M_J^{(0,k)}. \tag{4.8}$$

In the $k$-th Newton step of the algorithm *Newton-CE-MG* we shall apply the PROCEDURE MG $m$ times. For $i = 1, \ldots, m$, we take:

- $J$ from the discretization procedure, see e.g. (3.8).
- $H_j = \mathbb{R}^{N_j}$, $0 \leqslant j \leqslant J$, as in (4.6)—case $m = 1$.
- Mappings $\mathcal{A}_j$, $j = 0, \ldots, J$, defined as the discretization of the mapping $\mathcal{L}_i(u_0)$ on the grids $(T_j, B_j)$ by the method (3.6)–(3.7) with $u_0$ from step (N1) of the PROCEDURE NEWTON and $\mathcal{L}_i$ defined as in (3.13). (Note that here we must ensure symmetry and positive definiteness of the diagonal blocks of the original Jacobian only.)
- $\mathcal{I}_j, \mathcal{R}_j$ and $n(j)$ as in the algorithm *Newton-CG-MG*.

We denote the mappings $M_J$ from the PROCEDURE MG with each of these $m$ settings just described by $M_J^{(i)}$, $i = 1, \ldots, m$. Then we set

$$B^{CE} = \begin{pmatrix} M_J^{(1)} & 0 & \ldots & 0 \\ 0 & M_J^{(2)} & \ldots & 0 \\ \ldots & & & \ldots \\ 0 & 0 & \ldots & M_J^{(m)} \end{pmatrix} A_k^T. \tag{4.9}$$

(Recall our notation $A_k \equiv g'(u_k)$.) Note that for the solution of linear systems in the algorithm *Newton-CE-MG*, the multigrid method is applied to the same set of matrices in every Newton step.

In this section we state our two main convergence theorems, Theorem 5.1 and Theorem 5.2. Proofs of these two theorems are very similar, but in fact, different lemmas have to be utilized.

**Theorem 5.1.** (Convergence of the algorithm Newton-CG-MG.) *Consider the problem* (3.1) *and suppose that the assumptions* (A1)–(A5), (D1), (D2), *strong monotonicity condition*

$$(\exists C_0 > 0)(\forall x \in \Omega)(\forall \xi, \eta \in \mathbb{R}^{3m})$$

$$\sum_{i=1}^{m}\sum_{j=0}^{2}[a_{ij}(x;\xi) - a_{ij}(x;\eta)](\xi_{ij} - \eta_{ij}) > C_0 \sum_{i=1}^{m}\sum_{j=1}^{2}(\xi_{ij} - \eta_{ij})^2 \tag{5.1}$$

*and symmetry condition*

$$(\forall x \in \Omega)(\forall \xi \in \mathbb{R}^{3m})(\forall i,k,i,k = 1,\dots,m)(\forall j,l,j,l = 0,1,2)$$

$$\Big(\frac{\partial a_{ij}(x;\xi)}{\partial \xi_{kl}} = \frac{\partial a_{kl}(x;\xi)}{\partial \xi_{ij}}\Big) \tag{5.2}$$

*are valid. Let grids* $(T_j, B_j)$, $j = 0,\dots,J$, *on* $\Omega$ *be given. As shown above, the problem in the form* (3.2)–(3.4) *is well-defined and we can look for a solution of its approximation in the form* (3.6)–(3.8). *This leads to a system of algebraic equations in the form* (3.10):

$$g(u) = 0 \quad \textit{in} \quad \mathbb{R}^{mN_J}. \tag{5.3}$$

*Choose an arbitrary element* $u_0 \in \mathbb{R}^{mN_J}$. *If* $u_k$, $k \geqslant 0$, *are defined by the algorithm Newton-CG-MG applied to the system* (5.3) *and* $S_0$ *denotes the set*

$$S_0 = \{u \in H_J : \|g(u)\| \leqslant \|g(u_0)\|\}, \tag{5.4}$$

*then*

1. $(\forall k \geqslant 1)(u_k \in S_0)$, *the sequence of norms* $\|g(u_k)\|$ *is strictly decreasing and*

$$\lim_{k\to\infty} \|g(u_k)\| = 0.$$

2. $(\exists u^* \in S_0)(u^* = \lim_{k\to\infty} u_k) \wedge (g(u^*) = 0)$.

3. *Let*

$$\chi_k = \| \ (I - g'(u_k)L_k^{CG}) \ g(u_k) \ \| / \|g(u_k)\|. \tag{5.5}$$

*If*

$$\lim_{k\to\infty} \chi_k = 0 \tag{5.6}$$

*or*

$$(\exists r \in (0,1])(\forall k > k_0)(\chi_k \leqslant C_3 \|g(u_k)\|^r) \tag{5.7}$$

*is valid, then the convergence is superlinear or of the order* $r+1$*, respectively.*

**Theorem 5.2.** (Convergence of the algorithm Newton-CE-MG.) *Suppose that all the assumptions of Theorem* 5.1 *except the strong monotonicity condition* (5.1) *are valid. Again, we consider a system of algebraic equations in the form* (3.10) *obtained in the same way as in Theorem* 5.1,

$$g(u) = 0 \qquad in \quad \mathbb{R}^{mN_J}. \tag{5.8}$$

*Suppose that*

- $(\forall i, i = 1, \ldots, m)(\exists C_i > 0)(\dot{\forall} x \in \Omega)$

$$(\forall \xi, \eta \in \mathbb{R}^{3m} : (\forall j, k, j = 1, \ldots, m, k = 0, 1, 2)(j \neq i)(\xi_{jk} = \eta_{jk})$$

$$\left( \sum_{j=0}^{2} [a_{ij}(x;\xi) - a_{ij}(x;\eta)](\xi_{ij} - \eta_{ij}) > C_i \sum_{j=0}^{2} (\xi_{ij} - \eta_{ij})^2 \right), \tag{5.9}$$

- *in place of* (5.2) *only the following condition is valid*

$$(\dot{\forall} x \in \Omega)(\forall \xi \in \mathbb{R}^{3m})(\forall i, i = 1, \ldots, m)(\forall j, l, j, l = 0, 1, 2)$$

$$\left( \frac{\partial a_{ij}(x;\xi)}{\partial \xi_{il}} = \frac{\partial a_{il}(x;\xi)}{\partial \xi_{ij}} \right) \tag{5.10}$$

- *there is an element* $u_0 \in \mathbb{R}^{mN_J}$ *such that*

$$(\forall u \in \mathbb{R}^{mN_J} : \|g(u)\| \leqslant \|g(u_0)\|)(g'(u) \text{ is a regular mapping}). \tag{S1}$$

*Then, if $u_k$, $k \geqslant 0$, are defined by the algorithm Newton-CE-MG applied to the system* (5.8), *the assertions of Theorem* 5.1 *(with $L_k^{CE}$ instead of $L_k^{CG}$ in* (5.5), *of course) apply.*

Remark 5.1. Our proofs are based on Theorem 1 in Bank, Rose [2] stating that the above assertions are guaranteed by the following conditions:

(A-N1) The set $S_0 = \{u \in H_J \colon \|g(u)\| \leqslant \|g(u_0)\|\}$ is bounded.

(A-N2) The mapping $g$ is Fréchet differentiable and

$$(\forall u \in S_0)(g'(u) \text{ is regular and continuous}).$$

(A-N3) The mapping $L_k$ from (N3) satisfies

$$(\exists C_1 > 0)(\forall u \in S_0)(\forall k \in N)(\|L_k\| \leqslant C_1).$$

Denote $S_1 = \{u\colon \|u\| \leqslant \sup_{v \in S_0} \|v\| + C_1 \|g(u_0)\|\}$.

(A-N4) $(\exists C_2 > 0)(\forall u, v \in S_1)(\|g'(u) - g'(v)\| \leqslant C_2 \|u - v\|)$.

(A-N5) $\chi_0 \in (0,1)$ and $(\forall k \geqslant 1)(\chi_k \leqslant \chi_0)$, where $\chi_k$ is defined as in (5.5)

Note that, as discussed in [2, Section 3], other conditions mentioned in [2, Theorem 1] are satisfied automatically by the PROCEDURE NEWTON. But before starting to verify these conditions, we shall prove some lemmas concerning the algorithms *CG, CE* and *MG*:

**Lemma 5.1.** *Let the assumptions of Theorem* 5.1 *be valid. Then, if $u_k$, $k \geqslant 0$, are defined by the algorithm Newton-CG-MG applied to the system* (5.3), *the following is valid:*

$$(\forall k, k \leqslant 0)(M_J^{(0,k)} \textit{ is symmetric, positive definite}). \tag{5.11}$$

$$(\forall k, k \leqslant 0)(\exists \gamma_1^G, \gamma_2^G > 0)(\forall v \in \mathbb{R}^{mN_J})$$

$$(\gamma_1^G (A_k v, v) \leqslant (A_k M_J^{(0,k)} A_k v, v) \leqslant \gamma_2^G (A_k v, v)). \tag{5.12}$$

Proof. Assertion (5.11). It is easy to show that the conditions (5.1) and (5.2) ensure that

$$(\forall u \in \mathbb{R}^{mN_J})(g'(u) \text{ is symmetric, positive definite})$$

and thus the application of the PROCEDURE MG makes sense. Further, Theorem 5 in Bramble, Pasciak and Xu [3] states that if

$$\begin{aligned}&\text{the spectrum of the operator } (I - \mathcal{R}_j \mathcal{A}_j)(I - \mathcal{R}_j^T \mathcal{A}_j)\\ &\text{is in the interval } [0,1),\end{aligned} \tag{5.13}$$

then $M_J$ is symmetric, positive definite. In [3], see text near (3.4), Bramble, Pasciak and Xu also say that the condition (5.13) immediately follows from the condition

(A-MG1) $(\exists C_R > 0)(\forall j, j = 1,\dots, J)(\forall u \in H_j)$

$$\frac{\|u\|_j^2}{\lambda_j} \leqslant C_R(I - (I - \mathcal{R}_j\mathcal{A}_j)(I - \mathcal{R}_j^T\mathcal{A}_j)\mathcal{A}_j^{-1}u, u\ )_j, \tag{5.14}$$

where $\lambda_j$ is the greatest eigenvalue of the matrix $\mathcal{A}_j$ in question.

As is shown e.g. in Pospíšek [13], proof of Th.8.1, the settings of the algorithm $M_J^{(0,k)}$ are such that for all $k$, $k \geqslant 0$, the condition (A-MG1) is satisfied, so the assertion (5.11) holds. □

- Assertion (5.12). Theorem 6 in [3] states that the assertion (5.12) is valid if the conditions (A-MG1) and

(A-MG2) $(\exists \alpha, 0 < \alpha \leqslant 1)(\exists C_\alpha > 0)(\forall j, j = 1,\dots, J)(\forall u \in H_j)$

$$|(\mathcal{A}_j(I - \mathcal{I}_j\mathcal{P}_{j-1})u, u\ )_j| \leqslant C_\alpha\Big(\frac{\|\mathcal{A}_j u\|_j^2}{\lambda_j}\Big)^\alpha(\mathcal{A}_j u, u)_j^{1-\alpha} \tag{5.15}$$

are satisfied. For verification of the condition (A-MG2), again see e.g. Pospíšek [13], proof of Th. 8.1.

**Lemma 5.2.** *Let the assumptions of Theorem 5.2 be valid. Then for any $u_0 \in \mathbb{R}^{mN_J}$ the following holds:*

$$(\forall l, l = 1,\dots, m)(M_J^{(l)} \text{ is symmetric, positive definite}). \tag{5.16}$$

$$(\forall l, l = 1,\dots, m)(\exists \gamma_1^E, \gamma_2^E > 0)(\forall v \in \mathbb{R}^{N_J})$$

$$(\gamma_1^E(\mathcal{A}_J^{(i)}v, v) \leqslant (\mathcal{A}_J^{(i)}M_J^{(l)}\mathcal{A}_J^{(i)}v, v) \leqslant \gamma_2^E(\mathcal{A}_J^{(i)}v, v)), \tag{5.17}$$

*where $\mathcal{A}_J^{(i)}$ denotes the mapping $\mathcal{A}_J$ as used in the definition of the mapping $M_J^{(i)}$.*

Proof. Similarly as in the proof of Lemma 5.1, the conditions (5.9) and (5.10) imply that (for the meaning of $g_{ii}$, see (3.12))

$$(\forall i, 1 \leqslant i \leqslant m)(\forall u \in \mathbb{R}^{mN_J})(g'_{ii}(u) \text{ is symmetric, positive definite})$$

and thus the mappings $M_j^{(i)}$, $i = 1,\dots, m$, are well-defined. Now we can go, for $i = 1,\dots, m$, through the same steps as in the proof of Lemma 5.1 and complete the proof of Lemma 5.2. □

**Lemma 5.3.** *a) Let the assumptions of Theorem* 5.1 *be valid. If* $u_k$ *are defined by the algorithm Newton-CG-MG applied to the system* (5.3), *then*

$$(\forall f \in H_J)(\forall k \in N)(\forall u_k \in S_0)$$

$$(\exists q_{r,k} \in [0,1))(\|(L_k^{CG} - A_k^{-1})f\|_{A_k} \leqslant q_{r,k}\|A_k^{-1}f\|_{A_k}). \tag{5.18}$$

*b) Let the assumptions of Theorem* 5.2 *be valid. If* $u_k$ *are defined by the algorithm Newton-CE-MG applied to the system* (5.8) *and* $B_0 \equiv B^{CE}(A_k^T)^{-1}$ *then*

$$(\forall f \in H_J)(\forall k \in N)(\forall u_k \in S_0)$$

$$(\exists \overline{q}_{r,k} \in [0,1))(\|(L_k^{CE} - A_k^{-1})f\|_{B_0} \leqslant \overline{q}_{r,k}\|A_k^{-1}f\|_{B_0}). \tag{5.19}$$

Remark 5.2. Note that the algorithms in the PROCEDURE CG/CE start with zero initial approximation. In fact, Lemma 5.3 says that the algorithms proposed to solve the appropriate systems of linear equations are—under given conditions—always convergent.

Proof of Lemma 5.3. The proof is based on the convergence theorem for general two-step conjugate direction methods, see e.g. Samarskij, Nikolajev [17, p. 355], applied to special cases of the conjugate gradient method and the congujate error method.

In the case of the conjugate gradient method the theorem from [17] mentioned above says that if

$$A\colon H \to H,\ B^{CG}\colon H \to H \quad \text{are symmetric, positive definite,} \tag{5.20}$$

and if $x_{\varrho+1}$ are defined by the PROCEDURE CG from Section 4, then

$$\lim_{\varrho\to\infty} \|L^{CG}(A,b,\varrho) - A^{-1}b\| = 0$$

and

$$\|x_{\varrho+1} - A^{-1}b\|_A \leqslant q_r\|x_0 - A^{-1}b\|_A \tag{5.21}$$

with $q_r \in [0,1)$. Further, this $q_r$ can be expressed in terms of $\gamma_1$, $\gamma_2$ for which the following is valid:

$$(\exists \gamma_1,\gamma_2 > 0)(\forall v \in H)(\gamma_1(Av,v) \leqslant (AB^{CG}Av,v) \leqslant \gamma_2(Av,v)). \tag{5.22}$$

In the case of the conjugate error method the conditions (5.20) are replaced by

(5.23) $$A\colon H \to H \quad \text{is an arbitrary regular mapping,}$$

(5.24) $$B_0 \equiv B^{CE}(A^T)^{-1},\ B_0\colon H \to H \quad \text{is symmetric, positive definite}$$

and the estimate (5.21) is replaced by

(5.25) $$\|x_{\varrho+1} - A^{-1}b\|_{B_0} \leqslant \overline{q}_r \|x_0 - A^{-1}b\|_{B_0}$$

with $\overline{q}_r \in [0,1)$. Here $\overline{q}_r$ can also be expressed in terms of $\gamma_1$, $\gamma_2$ for which the following is valid:

(5.26) $$(\exists \gamma_1, \gamma_2 > 0)(\forall v \in H)(\gamma_1(B_0 v, v) \leqslant (A^T A v, v) \leqslant \gamma_2(B_0 v, v).)$$

The assertion a) now follows from Lemma 5.1 which ensures the validity of the conditions (5.20), and from the fact that $x_0 = 0$ in the PROCEDURE CG. Similarly, the assertion b) follows from Lemma 5.2 which ensures the validity of the conditions (5.23) and (5.24), and from setting $x_0 = 0$ in the PROCEDURE CE. □

**Lemma 5.4.** (Ortega, Rheinboldt [12, Th. 5.4.3, p. 142]) *Let $g\colon D \subset \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable mapping on an open convex subset $D_0 \subset D$. Then*

*$g$ is strongly monotone in $D_0$*

*iff*

$$(\exists \gamma > 0)(\forall u \in D_0)(\forall \xi \in \mathbb{R}^n)((g'(u)\xi, \xi) \geqslant \gamma(\xi, \xi)).$$

**Lemma 5.5.** (Samarskij, Nikolajev [17, Th. 2, p. 227]) *Let $A\colon \mathbb{R}^n \to \mathbb{R}^n$ be a symmetric positive definite mapping and let $\delta > 0$ be such that*

$$(\forall \xi \in \mathbb{R}^n)((A\xi, \xi) \geqslant \delta(\xi, \xi)).$$

*Then the norm of the mapping $A^{-1}$ inverse to $A$ can be estimated by*

$$\|A^{-1}\| \leqslant \delta^{-1}.$$

P r o o f of Theorem 5.1. We shall show that the conditions (A-N1)–(A-N5) are satisfied.

(A-N1) The coercivity of the mapping $A$ in (3.2) clearly implies that $g$ is coercive. Rearranging the well-known Schwarz inequality to the form

$$(g(v), v)/\|v\| \leqslant \|g(v)\| \quad (v \neq 0),$$

we see that $\|g(v)\| \to \infty$ as $\|v\| \to \infty$. Assume that the set $S_0$ in (A-N1) is not bounded. Then there exists a sequence

$$\{v_\nu : v_\nu \in S_0\}_{\nu \geqslant 1}, \quad \|v_\nu\| \to \infty$$

and thus, by coercivity, also $\|g(v_\nu)\| \to \infty$. But this is a contradiction with the definition of $S_0$.

(A-N2) Differentiability of $g(u)$ and continuity of $g'(u)$ on $S_0$ follow from smoothness of the functions $a_{ij}$ in (3.1). Regularity and even symmetry and positive definiteness of $g'(u)$ were already stated in Lemma 5.1.

(A-N3) We start with the triangle inequality

$$\|L_k^{CG}\| \leqslant \|L_k^{CG} - A_k^{-1}\| + \|A_k^{-1}\|. \tag{5.27}$$

The mapping $g$ is strongly monotone in $S_0$ and thus by Lemma 5.4

$$(\exists \gamma > 0)(\forall v \in S_0)(\forall \xi \in V_J)((g'(v)\xi, \xi) \geqslant \gamma \|\xi\|^2). \tag{5.28}$$

Hence

$$(\exists \gamma > 0)(\forall k \in \mathbb{N})(\forall \xi \in V_J)((A_k \xi, \xi) \geqslant \gamma \|\xi\|^2). \tag{5.29}$$

Then, by Lemma 5.5,

$$(\forall u \in S_0)(\|A_k^{-1}\| \leqslant \gamma^{-1}). \tag{5.30}$$

Now we shall show the uniform boundedness of $\|L_k^{CG} - A_k^{-1}\|$. From Lemma 5.3 we have

$$(\forall f \in V_J)(\forall k \in \mathbb{N})(\|(L_k^{CG} - A_k^{-1})f\|_{A_k} \leqslant q_{r,k}(\|A_k^{-1} f\|_{A_k}) \tag{5.31}$$

with $q_{r,k} \in [0,1)$. To estimate the left-hand side of this inequality from below, we use Lemma 5.4. We obtain (for $\gamma$ see in (5.29))

$$\begin{aligned} &(\forall f \in V_J)(\forall k \in \mathbb{N}) \\ &\quad (\|(L_k^{CG} - A_k^{-1})f\|_{A_k} \equiv (A_k(L_k^{CG} - A_k^{-1})f, (L_k^{CG} - A_k^{-1})f)^{1/2} \\ &\qquad \geqslant \gamma^{1/2} \|(L_k^{CG} - A_k^{-1})f\|). \end{aligned} \tag{5.32}$$

To estimate the right-hand side of the inequality (5.31) from above, we use Lemma 5.5 with $\gamma$ from (5.29) used instead of $\delta$. By Lemma 5.5, this constant does not depend on $k$, hence

$$(\forall f \in V_J)(\forall k \in \mathbb{N})$$

$$q_{r,k}\|A_k^{-1}f\|_{A_k} = q_{r,k}(f, A_k^{-1}f)^{1/2} < \|A_k^{-1}\|^{1/2}\|f\| \leqslant \gamma^{-1/2}\|f\|. \tag{5.33}$$

Substituting inequalities (5.32) and (5.33) into (5.31) gives

$$\|(L_k^{CG} - A_k^{-1})f\| < \gamma^{-1}\|f\|$$

and thus

$$(\forall f \in V_J)(\forall k \in \mathbb{N})\|L_k^{CG} - A_k^{-1}\| < \gamma^{-1}. \tag{5.34}$$

The assumption (A-N3) now follows from (5.27), (5.30) and (5.34).

(A-N4) Follows from the smoothness of the coefficients $a_{ij}$, as in (A-N2).

(A-N5) Note that the value of $\chi_k$ can be computed easily in practice:

$$\chi_k = \|g(u_k) + g'(u_k)v_k\|/\|g(u_k)\|,$$

where $v_k \equiv L^{CG}(-g(u_k))$ from the step (N3) of the algorithm Newton. For example, to obtain the convergence of $(p+1)$-st order, $p \in (0,1]$, we stop the algoritm *CG-MG* when

$$\chi_k \leqslant \chi_0(\|g(u_k)\|/\|g(u_0)\|)^p.$$

□

Proof of Theorem 5.2. We will show that the conditions (A-N1)–(A-N5) are satisfied.

(A-N1) The same as in the proof of Theorem 5.1.

(A-N2) Differentiability of $g(u)$ and continuity of $g'(u)$ on $S_0$ follow from smoothness of the functions $a_{ij}$ in (3.1), regularity of $g'(u)$ is the assumption of our theorem.

(A-N3) We start with the triangle inequality

$$\|L_k^{CE}\| \leqslant \|L_k^{CE} - A_k^{-1}\| + \|A_k^{-1}\|. \tag{5.35}$$

The mapping $D\colon v \mapsto \|g'(v)^{-1}\|$ is continuous for $v \in S_0$ and, due to the fact that $S_0$ is a bounded and closed set in a finite-dimensional space (and hence is compact), $D$ attains its maximum $C_A$ on $S_0$. We have

$$(\forall k \in \mathbb{N})(\forall u_k \in S_0)(\|A_k^{-1}\| \leqslant C_A). \tag{5.36}$$

Now, estimate the term $\|L_k^{CE} - A_k^{-1}\|$. Using equivalence of the norms $\|\cdot\|_{B_0}$ and $\|\cdot\|$, i.e.

$$(\exists\gamma_1 > 0)(\exists\gamma_2 > 0)(\forall\xi \in \mathbb{R}^{mN_J})(\gamma_1\|\xi\| \leqslant (B_0\xi,\xi)^{1/2} \leqslant \gamma_2\|\xi\|),$$

we obtain from (5.19)

$$\|(L_k^{CE} - A_k^{-1})f\| \leqslant \bar{q}_{r,k}\gamma_2/\gamma_1\|A_k^{-1}\|\|f\|^2.$$

Combining this with the inequality (5.36), we have

$$\|L_k^{CE}\| \leqslant C_A(1+\gamma_2/\gamma_1).$$

(A-N4), (A-N5) The same as in the proof of Theorem 5.1. □

## 6. Application to the semiconductor device equations

### 6.1. Model Problem.

In 1950, Van Roosbroeck [15] proposed a system of three coupled nonlinear partial differential equations as a basic mathematical model describing electro-physical behaviour of semiconductor devices. We will be interested in the following, rather simplified form of these equations, ignoring complications like variable mobilities, oxide regions and avalanche generation rate. Our problem, nonetheless, captures some of the difficulties that occur in practice and its satisfactory solution still represents a great challenge to numerical analysis:

$$\begin{aligned}
&(6.1) && -\operatorname{div}(\operatorname{grad} u) + \mathrm{e}^u\eta - \mathrm{e}^{-u}\nu = D_C, \\
&(6.2) && -\operatorname{div}(\mathrm{e}^u \operatorname{grad}\eta) + Q(u,\eta,\nu)(\eta\nu - 1) = 0, \qquad x \in \Omega, \\
&(6.3) && -\operatorname{div}(\mathrm{e}^{-u} \operatorname{grad}\nu) + Q(u,\eta,\nu)(\eta\nu - 1) = 0,
\end{aligned}$$

$$(6.4) \qquad u = u_D,\ \eta = \eta_D,\ \nu = \nu_D, \qquad x \in \Gamma_D,$$

$$(6.5) \qquad \frac{\partial u}{\partial \vec{n}} = \mathrm{e}^u\frac{\partial \eta}{\partial \vec{n}} = \mathrm{e}^{-u}\frac{\partial \nu}{\partial \vec{n}} = 0, \quad x \in \Gamma_N$$

where

$$(6.6) \qquad D_C \in L_\infty(\Omega), \quad Q \in C(\mathbb{R}^3) \quad \text{and} \quad (u_D,\eta_D,\nu_D) \in [L_\infty(\overline{\Omega}) \cap C^1(\overline{\Omega})]^3.$$

We will use the following notation:

$$U_D \equiv (u_D, \eta_D, \nu_D),\quad V^\infty \equiv [V^2 \cap L_\infty(\Omega)]^3,\quad W^\infty \equiv [W^{1,2}(\Omega) \cap L_\infty(\Omega)]^3.$$

**Definition 6.1.** Let, as in (2.1), $V = \prod_{i=1}^{m} V^{p_i}$, $1 < p_i < \infty$, $1 \leqslant i \leqslant m$, and let a mapping $A^S\colon V^\infty \to V^*$ and a functional $f^S \in V^*$ be defined as follows:

$$\begin{aligned}
&(\forall U \in V^\infty, U \equiv (u,\eta,\nu))(\forall \Phi \in V, \Phi \equiv (\varphi_1,\varphi_2,\varphi_3)),\\
&\Big(\langle A^S U, \Phi\rangle_V = \sum_{i=1}^{3}\sum_{j=0}^{2}\int_\Omega \operatorname{grad} u \operatorname{grad} \varphi_1\\
&\qquad + \mathrm{e}^{u} \operatorname{grad}\eta \operatorname{grad}\varphi_2 + \mathrm{e}^{-u} \operatorname{grad}\nu \operatorname{grad}\varphi_3\\
&\qquad + (\mathrm{e}^{u}\eta - \mathrm{e}^{-u}\nu)\varphi_1 + Q(u,\eta,\nu)(\eta\nu - 1)(\varphi_2+\varphi_3)\,\mathrm{d}x\Big),
\end{aligned}$$

$$(\forall \Phi \in V)\quad \Big(\langle f^S, \Phi\rangle_V = \int_\Omega D_C(x)\varphi_1(x)\,\mathrm{d}x\Big).$$

We say that $U_S = (u,\eta,\nu) \in W^\infty$ *is a solution of the problem* (6.1)–(6.5) *in the space* $W^\infty$, if

$$U_S = U_S^* + U_D, \tag{6.7}$$

where $U_S^* \in V^\infty$ and

$$A^S U_S^* = f^S \text{ in } V^*. \tag{6.8}$$

As is shown in Pospíšek [14], we can consider another, regularized problem, solutions of which are also solutions of the problem (6.1)–(6.5) in the space $W^\infty$. This problem reads as follows:

$$- \operatorname{div}(\operatorname{grad} u) + \mathrm{e}^{P_E u} P_{GH}\eta - \mathrm{e}^{P_E(-u)} P_{GH}\nu = D_C, \tag{6.9}$$

$$- \operatorname{div}(\mathrm{e}^{P_E u} \operatorname{grad}\eta) + Q(P_E u, P_{GH}\eta, P_{GH}\nu)(P_{GH}\eta P_{GH}\nu - 1) = 0,\quad x \in \Omega, \tag{6.10}$$

$$- \operatorname{div}(\mathrm{e}^{P_E(-u)} \operatorname{grad}\nu) + Q(P_E u, P_{GH}\eta, P_{GH}\nu)(P_{GH}\eta P_{GH}\nu - 1) = 0, \tag{6.11}$$

$$u = u_D,\ \eta = \eta_D,\ \nu = \nu_D,\qquad x \in \Gamma_D, \tag{6.12}$$

$$\frac{\partial u}{\partial \vec n} = \mathrm{e}^{P_E u}\frac{\partial \eta}{\partial \vec n} = \mathrm{e}^{P_E(-u)}\frac{\partial \nu}{\partial \vec n} = 0,\qquad x \in \Gamma_N, \tag{6.13}$$

where $E$, $G$, $H$ are properly chosen constants, $P_E \equiv P_{-EE}$ and

$$(P_{rs}g)(x) = \begin{cases} r & \text{if } g(x) \leqslant r, \\ g(x) & \text{if } r < g(x) < s, \\ s & \text{if } s \leqslant g(x) \end{cases}$$

for any real function $g$.

Clearly, the problem (6.9)–(6.13) is in the form (3.1) where

$$(6.14)\quad a_{1j}(x;\xi) = \begin{cases} e^{P_E\xi_{10}} P_{GH}\xi_{20} - e^{P_E(-\xi_{10})} P_{GH}\xi_{30}, & j = 0, \\ \xi_{1j}, & j = 1,2, \end{cases}$$

$$(6.15)\quad a_{2j}(x;\xi) = \begin{cases} Q(P_E\xi_{10}, P_{GH}\xi_{20}, P_{GH}\xi_{30})(P_{GH}\xi_{20}P_{GH}\xi_{30} - 1), & j = 0, \\ e^{P_E\xi_{10}}\xi_{2j}, & j = 1,2, \end{cases}$$

$$(6.16)\quad a_{3j}(x;\xi) = \begin{cases} Q(P_E\xi_{10}, P_{GH}\xi_{20}, P_{GH}\xi_{30})(P_{GH}\xi_{20}P_{GH}\xi_{30} - 1), & j = 0, \\ e^{P_E(-\xi_{10})}\xi_{3j}, & j = 1,2, \end{cases}$$

$$(6.17)\quad \begin{gathered} f_1 = D_C,\ f_2 = f_3 = 0, \\ d_1 = u_D,\ d_2 = \eta_D,\ d_3 = \nu_D,\ h_1 = h_2 = h_3 = 0. \end{gathered}$$

Suppose first that $\eta$ and $\nu$ are known. Then only the equation (6.9) with its boundary conditions remains to be solved. It is easy to see that Theorem 5.1 can be applied to such a problem and hence the algorithm *Newton-CG-MG* can be used to solve the appropriate sets of algebraic equations. One can also verify that Theorem 5.1 is applicable in the cases when the pairs $u$, $\eta$ and $u$, $\nu$ are supposed to be known, and use this fact in solving the problem (6.9)–(6.13) by the nonlinear block Gauss-Seidel method (see e.g. Ortega, Rheinboldt [12]) with the blocks being defined by subdividing the original system into three sets of equations corresponding to (6.9)–(6.11). If the partial differential equations in (6.9)–(6.11) are only weakly coupled, the method is very effective. Moreover, standard procedures and (fast) algorithms for elliptic type problems with potential operators (like the algorithm *Newton-CG-MG*) can be used to solve the appropriate sets of equations. However, convergence theorems are restricted to only a few special cases (Jerome [9], Kerkhoven [10]) and the nonlinear block Gauss-Seidel algorithm seems not to be convergent for many other practically important situations.

In this paper, a procedure for the solution of the problem (6.9)–(6.13) which is based on the algorithm *Newton-CE-MG* is proposed and summarized in the next theorem.

**Theorem 6.1.** *Consider the problem* (6.9)–(6.13). *Suppose that for an integer* $J \geqslant 0$ *a sequence of grids* $(T_j, B_j)$, $j = 0, \ldots, J$ *is given as in Section* 3, *such that*

*$T_0$ is of acute type. Let us define a weak solution of the problem* (6.9)–(6.13) *and discretize the associated problem on the grids* $(T_J, B_J)$ *by the method* (3.6)–(3.7). *We obtain a system of nonlinear algebraic equations in the form* (3.10)*:*

$$g(u) = 0 \quad \text{in } \mathbb{R}^{mN_J}. \tag{6.18}$$

*Suppose further that we have an element $u_0 \in \mathbb{R}^{3N_J}$ such that the condition* (S1) *is valid. Then, if $u_k$, $k \geqslant 1$, are defined by the algorithm Newton-CE-MG applied to the system* (6.18), *the assertions of Theorem* 5.2 *apply.*

Proof. In Pospíšek [14], validity of assumptions (A1), (A2), ..., (A5), (D1) and (D2) was proved. Verification of the remaining assumptions of Theorem 5.2—i.e. (5.9) and (5.10)—is simple. Thus, the assertions of Theorem 5.2 apply. □

## 7. Conclusion

In practice, either Gaussian elimination, or iteration schemes based on various generalizations of the conjugate gradient method (BiCG [5], CGS [19]) and the conjugate residual method (ORTHOMIN [20], GMRES [16]) are used instead of the procedure CE-MG proposed in this paper. However, the resulting algorithm is then very slow, or its convergence theory is available in some special cases only. On the other hand, this paper together with [14] represents a method of treating boundary value problems in the form (3.1), starting with their weak formulation, up to a convergent solution algorithm.

### *References*


[1] *R.E. Bank, H.D. Mittelmann*: Continuation and multi-grid for nonlinear elliptic systems. Multigrid Methods. Proceedings (Hackbusch, W., ed.). Lect. Notes Math., Berlin, Heilderberg, New York, 1985.

[2] *R.E. Bank, D.J. Rose*: Global approximate Newton methods. Numer. Math. *37* (1981), 279–295.

[3] *J.H. Bramble, J.E. Pasciak, J. Xu*: The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms. Math. Comp. *56* (1991), 1–34.

[4] *G. Brussino, V. Sonnad*: A comparison of direct and preconditioned iterative techniques for sparse, unsymmetric systems of linear equations. Int. J. Numer. Meth. Eng. *28* (1989), 801–815.

[5] *R. Fletcher*: Conjugate gradient methods for indefinite systems. Proc. Dundee Biennal Conference on Numerical Analysis (Watson, G.A., ed.). Springer-Verlag, New York, 1975.

[6] *R. Fletcher*: Multigrid FAS methods for the solution of system of nonlinear partial differential equations occuring in semiconductor device simulation. Third multigrid seminar (Biesenthal 1988). Rep. Math. 89-03, Akad. Wiss. DDR, Berlin, 1989, pp. 1–10.

[7] *H.K. Gummel*: A self-consistent iterative scheme for one-dimensional steady state transistor calculations. IEEE Trans. on Electron Devices, ED-11 (1964), 455–465.

[8] *W. Hackbusch*: Multigrid Methods and Applications. Springer, Berlin, Heidelberg, New York, Tokyo, 1985.

[9] *J.W. Jerome*: The role of semiconductor device diameter and energy-band bending in convergence of Picard iteration for Gummel map. IEEE Trans. on Computer-Aided Design, CAD-4, *4* (1985), 489–495.

[10] *T. Kerkhoven*: A spectral analysis of the decoupling algorithm for semicondutor simulation. SIAM J. Numer. Anal. *25* (1988), 1299–1312.

[11] *M.S. Mock*: Analysis of Mathematical Models of Semiconductor Devices. Dublin, Boole Press, 1983.

[12] *J.M Ortega, W.C. Rheinboldt*: Iterative Solution of Nonlinear Equations in Several Variables. New York, London, Academic Press, 1970.

[13] *M. Pospíšek*: Mathematical Methods in Semiconductor Device Modelling. PhD Thesis. Prague, MÚ ČSAV, 1991. (In Czech.)

[14] *M. Pospíšek*: Nonlinear boundary value problems with application to the semiconductor device equations. Appl. Math. *39* (1994), 241–258.

[15] *W.V. Van Roosbroeck*: Theory of flow of electrons and holes in germanium and other semiconductors. Bell Syst. Tech. J. *29* (1950), 560–607.

[16] *Y. Saad, M.H. Schultz*: GMRES: A generalized minimal residual algorithm for solving non-symmetric linear systems. Research Report 254. Department of Computer Science. Yale University. New Haven Conecticut, 1983.

[17] *A.A. Samarskij, J.S. Nikolajev*: Solution Methods for Grid Equations. Praha, Academia, 1985. (In Czech.)

[18] *D.L. Scharfetter, H.K. Gummel*: Large signal analysis of a silicon Read diode oscillator. IEEE Trans. Electron Devices, ED-16 (1969), 64–77.

[19] *P. Sonneveld, P. Wesseling, P.M. de Zeeuw*: Multigrid and conjugate gradient methods as convergence acceleration techniques. Multigrid Methods for Integral and Differential Equations (Paddon, D.J., Holstein, H., eds.). Clarendon Press, Oxford, 1985, pp. 117–167.

[20] *P.K.W. Vinsome*: ORTHOMIN, an iterative method for solving sparse sets of simultaneous linear equations. Paper SPE 5729. Society of Petroleum Engineers of AIME, 1976.

[21] *A.S. Shieh*: On the solution of coupled system of PDE by a multigrid method. IEEE Trans. on Computer-Aided Design, CAD-4 (1985), no. 4, 527–530.

*Author's address*: *Miroslav Pospíšek*, T-SOFT, spol. s r.o., Novodvorská 994, 140 00 Praha 4, e-mail: `pospisek@tsoft.cz`.